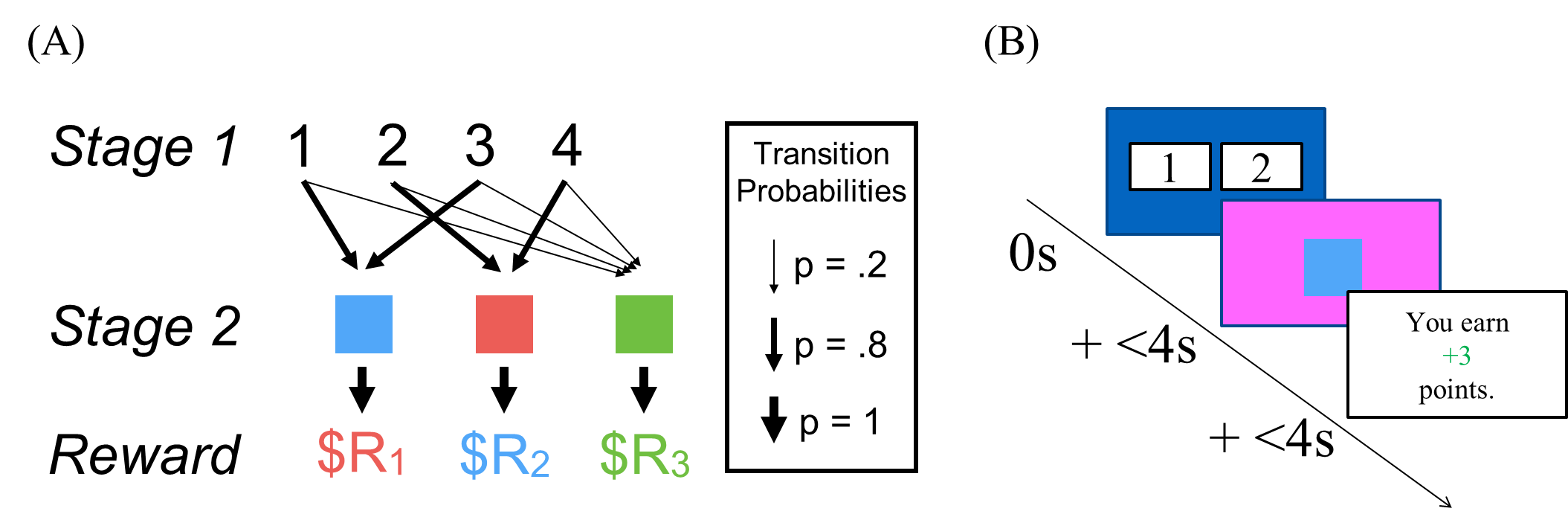
**Supplementary Materials**

**Experiment 1**

Behavioral Task

232 subjects were recruited on Amazon Mechanical Turk to participate in a two-stage Markov decision task. Stage 1 had four options (represented by the numbers 1 through 4), which each led probabilistically to one of three states (represented by the colors red, blue, and green). These states in turn had only one available action, which deterministically led to a reward. (See Figure 1.)

Figure 1: (A) Design of Experiment 1. Four number options lead probabilistically to one of three colors, which in turn lead to rewards. (B) Flow of the task. Subject is presented with two numbers, clicks on number 1, transitions to the blue square, clicks on it, and receives a reward.



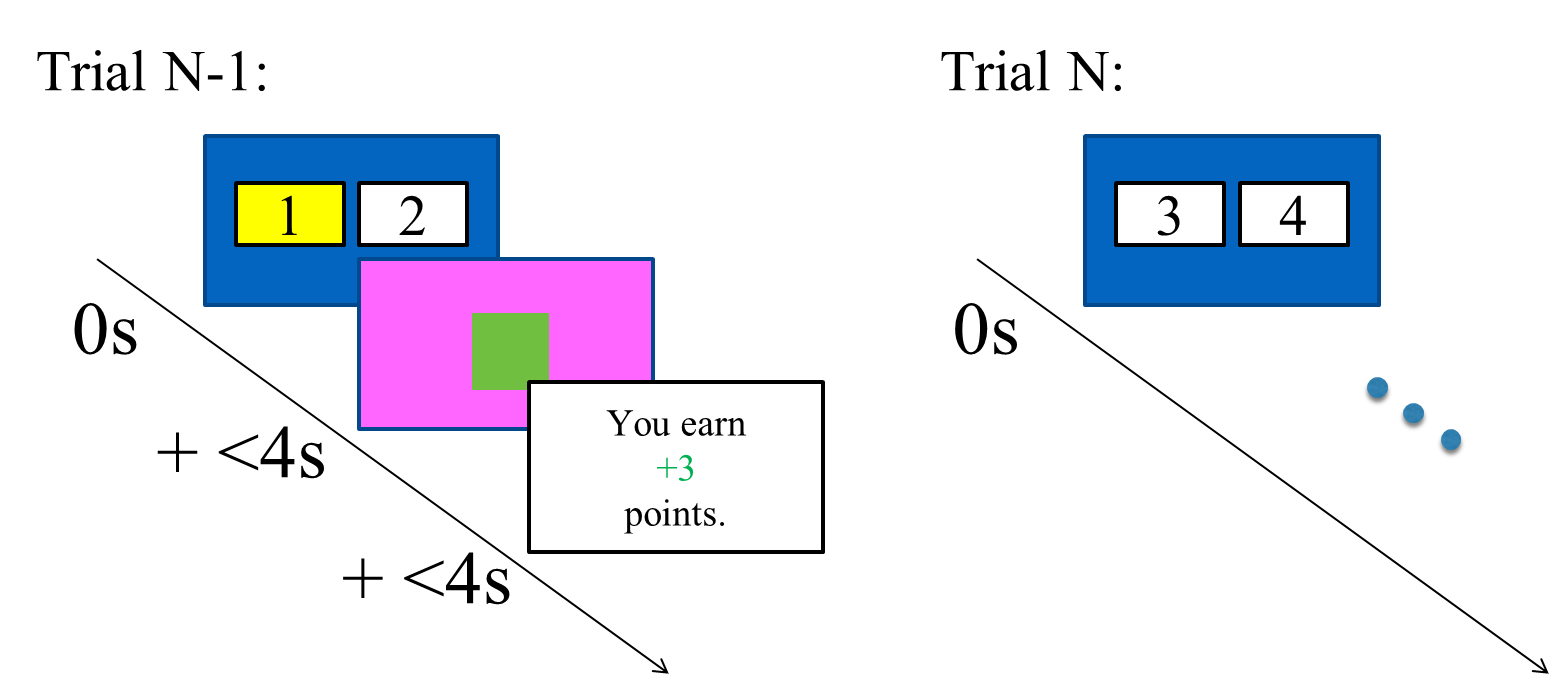
Numbers 1 and 3 each had an 80% chance of going to the blue square, and a 20% chance of going to the green square. 2 and 4 each had an 80% chance of going to the red square, and a 20% chance of going to the green square. Subjects were explicitly told these transition probabilities, and were trained on them in the practice rounds. The high-probability transition of each number became the “goal” of that number – the goal of clicking on 1 would be to get blue, the goal of clicking on 2 would be to get red, etc.

In each trial, subjects were presented with only two of the four number options. The option pairs to present were chosen randomly, with the constraint that the high-probability transitions of the two options had to lead to different colors – i.e. 1 and 3 could not be paired.

After clicking on one of the two numbers, subjects transitioned to a color, clicked on the color, and received a reward. The transitions from number to color were determined randomly according to the transition probabilities above. The rewards for each color were initialized uniformly at random on a range of -4 points to +5 points, and went on a bounded Gaussian random walk for the rest of the experiment. After each round, the drift was sampled from a normal distribution with (μ=0, σ=1.8), rounded to the nearest integer, and added to the current reward level[[1]](#footnote-1).

There was a critical subset of trials which we will call *congruent goal* trials. These trials had two defining characteristics: they immediately followed a low-probability transition to green, and they did not have the number that was chosen last round available as an option. Instead, they had the other number option with the same goal. For example, if the subject was presented with options 1 and 2, chose option 1, and transitioned to green, the next trial would be a congruent goal trial if it presented option 3 (paired with either 2 or 4). (See Figure 2.)

Figure 2: Congruent goal trial on trial N. On trial N-1, subject chooses option 1, transitions to green, and earns +3 points. The model-free-goal value is +3. On trial N, the subject is presented with options 3 and 4. (Critically, option 3 has the same goal as option 1.)



The reward received on the low-probability transition preceding a congruent goal trial will be called the *model-free-goal* value. This value is critical because neither a pure model-based system nor pure model-free system could incorporate it in a way that would influence decision-making on the congruent goal trial. A model-based system knows that green is equally reachable from any of the four number options, so forward planning is useless. A model-free system would only learn about the action which it took – in our example, selecting number 1 – and would not have learned anything about numbers 3 or 4. For a system to incorporate the reward, it would have to perform model-free learning on goal selection.

Subjects had 75 practice trials and 175 real trials. On the real trials, subjects had only 4 seconds to make their choice between the two numbers. If they did not make a choice within 4s, the trial would time out and the next trial would begin. Subjects were excluded from analysis if they timed out on more than 50 trials.

Analysis

We restricted our primary analyses to congruent goal trials.The independent variable in all the analyses was the model-free-goal value. The dependent variable was the subject’s choice on the congruent goal trial, coded as 1 if the chosen action had the same goal as the previous round and 0 otherwise.

We hypothesized that the model-free-goal value would predict subjects’ choices on the subsequent congruent goal trials. We tested this hypothesis in three ways. First, as a coarse test, we computed the mean choice when the model-free-goal value was less than 0 and the mean choice when it was greater than 0, and compared the two using a paired t-test. Second, as a more granular test, we regressed choice on reward using a logistic mixed-effects model.

Third, to definitively rule out any influence from a pure model-based or model-free system, we estimated a second mixed-effects model with approximate model-based and model-free action values as additional regressors. These values were calculated as follows. The model-based value of an action with a certain color goal was defined as the last reward which the subject received from that color, discounted over time (with discount parameter ). The model-free value of an action was defined as the reward received the last time the subject selected that action, discounted over time[[2]](#footnote-2).

These two values, model-based and model-free, were computed for both available action options in each critical trial. Then the model-based value of the action which the subject did not chose was subtracted from the model-based value of the action which the subject did chose (in accordance with the coding scheme of the dependent variable), and the resulting single value became the model-based regressor in the mixed-effects model. The same procedure was applied to the model-free values. Therefore, the second mixed-effect model had three regressors: model-based, model-free, and model-free-goal[[3]](#footnote-3).

To test the significance of the model-free-goal regressor in the mixed-effects models, we used the Wald test. We also estimated null models (which in each case was the full model with the model-free-goal regressor removed), and performed both likelihood ratio tests and parametric bootstrap analyses to assess whether the model-free-goal regressor increased the model’s likelihood enough to be justifiably included.

Finally, to benchmark the size of the model-free-goal effect, we also considered trials which followed low-probability transitions to green but were not critical trials, because the option chosen in the trial with the low-probability transition was available again. We call these *congruent goal+action* trials.

On congruent goal+action trials, the effect of the reinforcement on the low-probability transition could be incorporated by both a model-free and a model-free-goal system. Therefore, subtracting the effect size of the reinforcement on the congruent goal trials from the effect size on these trials could give us a rough estimate of the size of the model-free effect alone. We would then be able to compare that magnitude to the magnitude of the model-free-goal effect. To measure the effect size of reinforcement on congruent goal+action trials, we ran both the paired t-test and the simpler mixed-effect model on them.

All analyses were conducted in R (cite).

Results

14 subjects were excluded for timing out on more than 50 trials, leaving 218 subjects and 6120 congruent goal trials for all the analyses. When the model-free-goal was positive, subjects chose the congruent goal 85.2% of the time. When it was negative, subjects chose the congruent goal 68.7% of the time. This difference was significant (paired t-test, t(217) = -11.2, p < .0001).

In the simple mixed-effects model, with only the model-free-goal reward as a regressor, the coefficient of the reward was .152 (Wald test, z = 11.5, p < .0001). The model was preferred to a null model without the reward (Likelihood ratio test, (2) = 312.1, p < .0001). In a parametric bootstrap analysis, 0 out of 1000 randomly resampled null models had a likelihood as large as the full model (p < .001).

In the complete mixed-effects model, with the model-based, model-free, and model-free-goal values as regressors, the coefficient of the model-free-goal value was .157 (z = 11.5, p < .0001). The model was preferred to the null model ((4) = 326.6, p < .0001). In a bootstrap analysis, 0 out of 1000 randomly resampled null models had a likelihood as large as the full model (p < .001). For comparison, the model-based coefficient was .152 (z = 6.37, p < .0001), and the model-free coefficient was .071 (z = 3.26, p < .005).

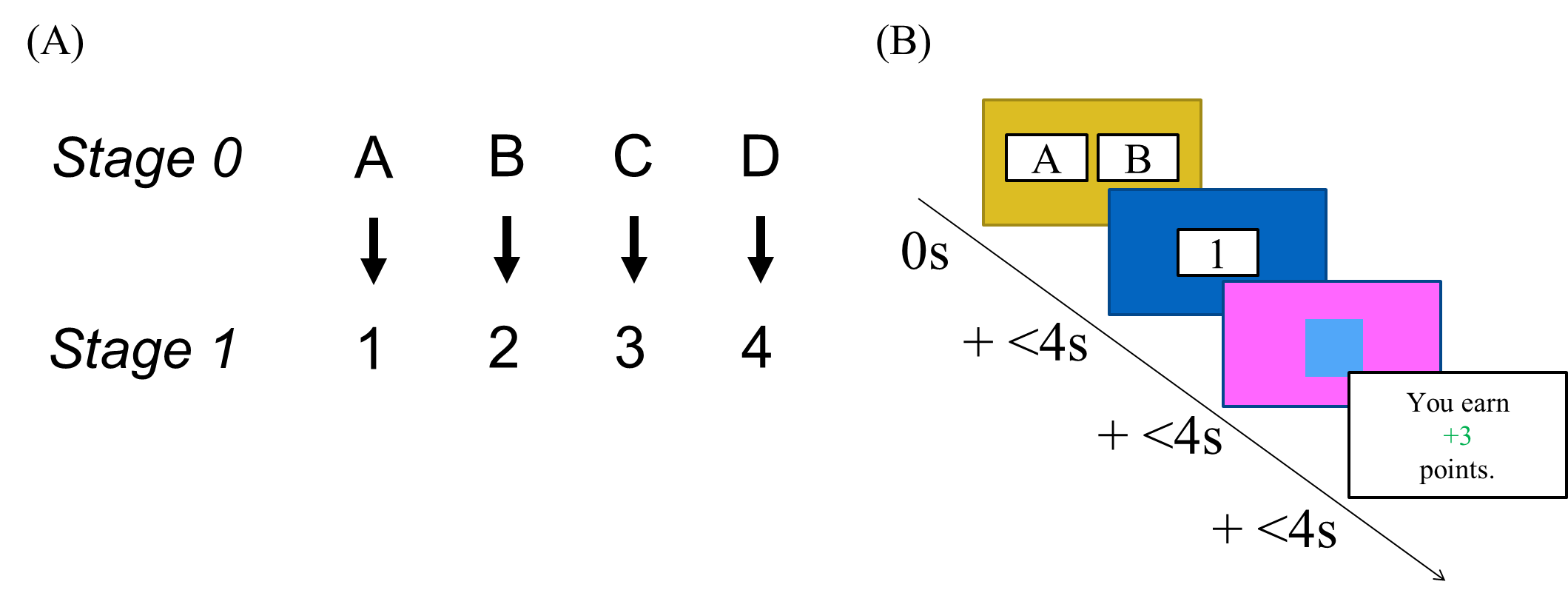
There were a total of 539 congruent goal+action trials. On these trials, subjects chose the same action 83.4% of the time after a reward and 72.3% of the time after a punishment (t(117) = -2.47, p < .02). In the simple mixed-effects model, the model-free-goal coefficient was .188. In the t-test, the effect size for congruent goal+action trials was slightly smaller than for congruent goal trials; in the mixed-effects model, it was slightly bigger. Thus we don’t know

**Experiment 2**

Behavioral Task

312 subjects were recruited online through Amazon Mechanical Turk. They performed a task identical to the one above, with one change. Before being exposed to the structure of the main task, subjects were trained on a set of intuitive, deterministic transitions from letters to numbers (Figure 3A).

Figure 3: (A) Transitions from Stage 0 letters to Stage 1 numbers. (B) Flow of congruent goal trials in Experiment 3. On these trials only, subjects choose between two letters and transition to a number, a color, and then a reward.



After becoming familiar with those transitions, subjects proceeded with the same task as above. All non-critical trials had exactly the same structure as in Experiment 1, with a choice between two numbers leading to a color, which in turn led to one of three drifting reward distributions. However, on congruent goal trials, subjects instead were presented with a choice between two letters. Subjects chose a letter and received a number (in accordance with the deterministic transitions in Figure 3A). They then clicked on that number and, in the usual way, got to a color which led them to a reward. The congruent goal trials thus required a goal-directed system to plan one extra step ahead (Figure 3B).

Analysis

All analyses were identical to those in Experiment 1.

Results

19 subjects were excluded, leaving 293 subjects and 8086 congruent goal trials for all analyses. After a reward, subjects chose the congruent goal 81.9% of the time. After a punishment, subjects chose the congruent goal 67.7% of the time. This difference was significant (t(292) = -10.9, p < .0001).

In the simple mixed-effects model, the model-free-goal coefficient was .118 (z = 11.3, p < .0001), and the model was preferred to a null model ((2) = 291.6, p < .0001; by bootstrapping, p < .001).

In the complete mixed-effects model, the model-free-goal coefficient was .118 (z = 11.3, p < .0001), and the model was preferred to a null model ((4) = 291.6, p < .0001; by bootstrapping, p < .001). For comparison, the model-based coefficient was .125 (z = 6.92, p < .0001) and the model-free coefficient was .055 (z = 2.92, p < .005).

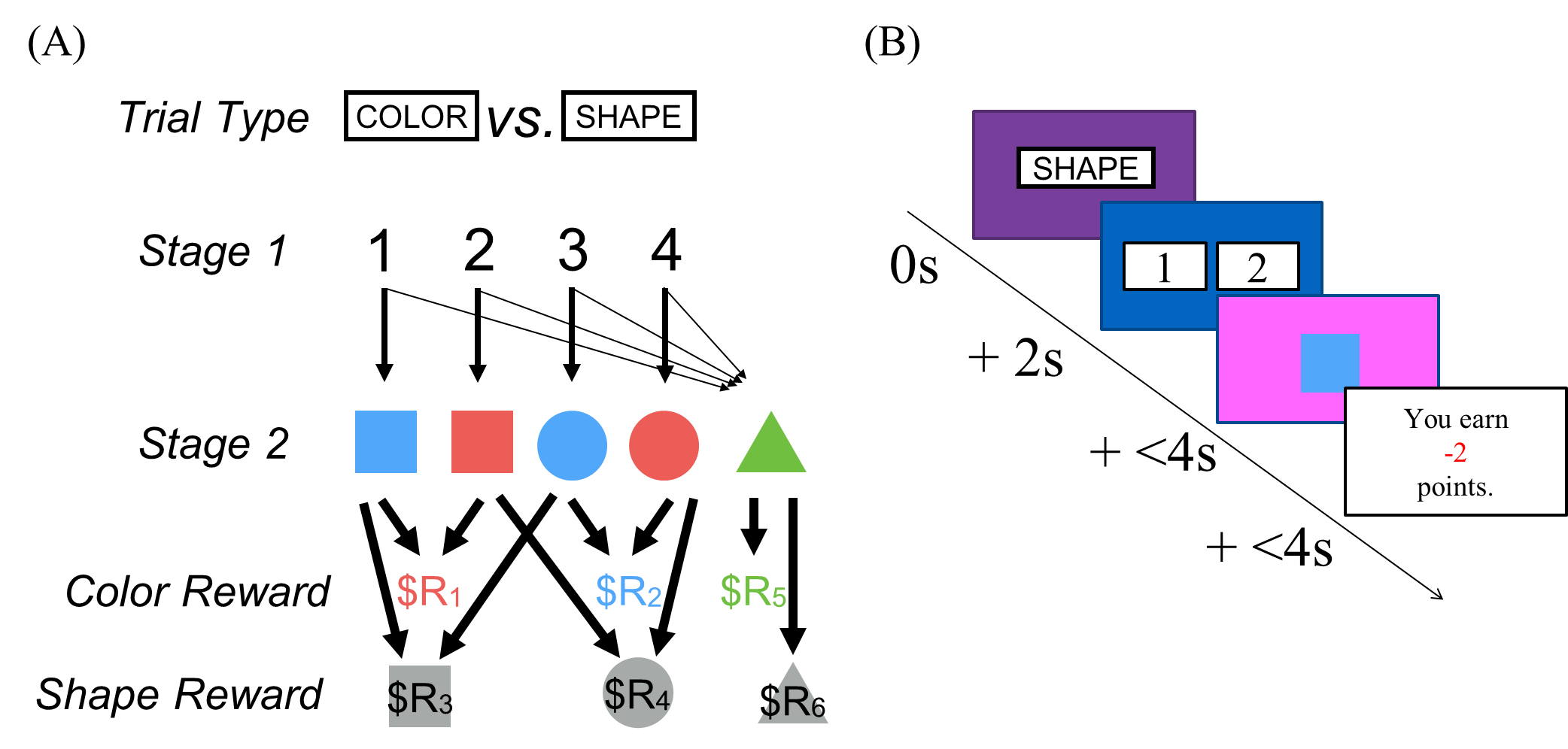
There were a total of 677 congruent goal+action trials. On these trials, subjects chose the same action 84.8% of the time after a reward and 70.2% of the time after a punishment (t(117) = -3.68, p < .001). In the simple mixed-effects model, the model-free-goal coefficient was .197. Here,

**Experiment 3**

Behavioral Task

191 subjects were recruited online through Amazon Mechanical Turk. They performed a task identical to the one in Experiment 1, with the following change. In Experiment 1, Stage 2 states only varied in their color (blue, red, or green). In Experiment 3, they also varied in their shape. There were three shapes: square, circle, and triangle. See Figure 4.

Figure 4: (A) Design of Experiment 3. Objects now differ in terms of color and shape. On color trials, the object color determines the reward; on shape trials, the shape does. (B) Task flow of Experiment 3. Subjects are told the trial type before proceeding with the trial.

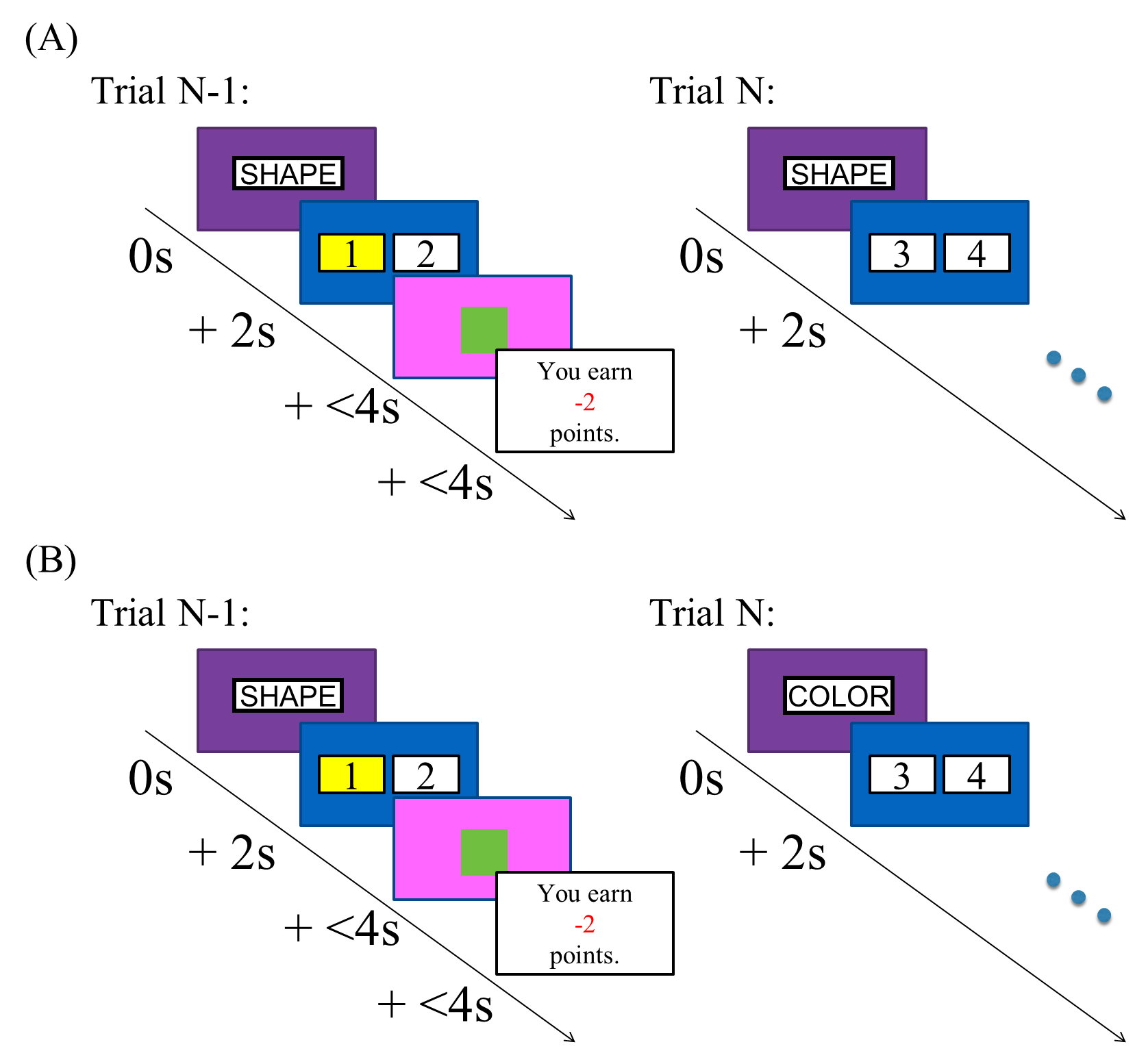


In Experiment 1, the color of an object determined the reward associated with it. On any given trial, blue had a certain reward, red had a certain reward, etc. In Experiment 3, each color and shape had a separate drifting reward distribution, and the reward value of an object could either be determined by its color or shape.

There were two trial types. On color trials, it was the color of the object which determined the reward. On shape trials, it was the shape of the object. Before each trial began, subjects were told the trial type. The flow of Experiment 3 is depicted in Figure 4B.

In Experiment 3, there were also two types of critical trials. In congruent goal trials, the trial type was the same as in the previous round. In incongruent goal trials, the two trial types were different[[4]](#footnote-4). (See Figure 5.)

Figure 5: (A) A congruent goal trial in Experiment 3. The trial types in the same in trials N-1 and N. (B) An incongruent goal trial. The trial types in N-1 and N are different.



Analysis

Our hypothesis predicts that the reward on the previous trial should predict subjects’ choices in congruent goal trials, not in incongruent goal trials. Therefore, we conducted the usual analyses on both congruent and incongruent trials separately, and then combined them into a new mixed-effects model with an interaction term between the model-free-goal regressor and the type of critical trial[[5]](#footnote-5).

Results

There were a total of 2473 congruent goal trials and 1254 incongruent goal trials over 176 subjects. On congruent goal trials, subjects chose the congruent goal 78.5% of the time after a reward and 75.0% of the time after a punishment. This difference just missed significance (t(175) = -1.9, p = .058). On incongruent goal trials, subjects chose what would have been the congruent goal 53.0% of the time after a reward and 53.6% of the time after a punishment. The difference was insignificant (t(166) = .196, p = .845).

In the simple mixed-effect model on the congruent goal trials, the model-free-goal coefficient was .039 (z = 2.68, p < .01). The model was preferred to a null model ((2) = 7.71, p < .025; by bootstrapping, p < .001).

In the complete mixed-effect model on congruent goal trials, the model-free-goal coefficient was .045 (z = 2.89, p < .005). Using a likelihood ratio test, the model was not preferred to a null model ((4) = 9.24, p = .055), but a bootstrap analysis showed that 0 out of 1000 randomly resampled null models had a likelihood as large as the full model (p < .001). For comparison, the model-based coefficient was 0.25485 (z = 6.528, p < .0001) and the model-free coefficient was .085 (z = 2.76, p < .01).

In the simple mixed-effect model on the incongruent goal trials, the coefficient was -.011 (z = -.716, p = .474). The model was not preferred to a null model ((2) = .512, p = .774; by bootstrapping, p = .75). We also estimated a model with both congruent and incongruent goal trials, which included the model-free-goal value and an interaction between that value and the trial type. In that model, the interaction term had a coefficient of .074 (z = 3.35, p < .001), and the model was preferred to a null model with the interaction term removed ((3) = 16.3, p < .001; by bootstrapping, p < .001). Congruent goal trials were coded as 1 and incongruent goal trials were coded as 0, so the positive interaction term indicates that the model-free-goal effect was significantly stronger for congruent goal trials.

There were a total of 204 congruent action trials. Subjects chose the same action 87.9% of the time after a reward and 77.6% of the time after a punishment. This difference was not significant (t(28) = -1.06, p = .3), and the effect size was ???. In the simple mixed-effect model, the coefficient for model-free-goal was .192. ???

**Simulations**

To validate our experiments, we built a generative model of agents who used some combination of model-based, model-free, and model-free-goal learning. We ran the agents through the same game as in Experiment 3, and showed that we would detect our result if and only if the agents were actually using model-free-goal learning.

Environment

The game was implemented as a Markov decision process with six states, the initial Stage 1 state and then one state for each Stage 2 object. State 1 had four possible actions (i.e. the four numbers), only two of which were available on any given trial. States 2-6 had only one possible action (i.e. clicking on the object), which led to a reward. The rewards were randomly generated for each agent by the same process as in the behavioral tasks.

Learning Mechanisms

The agents had three learning mechanisms. Their model-free reinforcement learning mechanism was the SARSA algorithm with eligibility traces (cite). Agents estimated a model-free value of the state-action pair (*s,a*), denoted *MFV(s,a)*. In stage 1, agents chose an action *a* and transitioned to state *s.* Their model-free update was:

In stage 2, agents chose the only available action (i.e. clicking on the object) and received reward *r*. Their model-free updates were:

was the learning rate, and was the eligibility trace. (Parameter selection is described below.)

Agents’ model-based learning mechanism was a type of dynamic programming (cite?). Agents maintained a model-based value of each state-action pair, denoted *MVB(s,a)*, but the model-based mechanism had a different conception of states than the model-free system. To the model-free system, each object was a different state. But because the model-based system knew that rewards were tied to object features (color or shape), not objects themselves, it conceptualized each object *feature* as a different state. The “feature-state” corresponding to a given object-state *s* is given by *.*

To calculate the model-based value of each action from state 1, agents estimated the transition probabilities of action *a* to state *s*, denoted by , by dividing the number of observed transitions from *a* to *s* by the total number of times *a* was selected. Then:

After transitioning to state *s*, clicking on the color, and receiving reward *r*, the model-based update was:

The third learning mechanism was our proposed mechanism, model-free learning on goal selection. After a trial with chosen action *a* and received reward *r*, agents inferred the intended goal *g(a)* by:

Agents then calculated a model-free value of each possible feature goal, denoted *MFG(g)*, by:

To determine the probability of an action *a*, agents took a weighted average *Wa* of the three values and entered it into a softmax function:

So agents were characterized by five parameters: (the learning rate), (the eligibility trace), (the softmax temperature), (the model-based weight), and (the model-free weight).

Simulation Process

In each simulation, 200 agents were generated with randomly sampled parameters[[6]](#footnote-6). We ran two simulations: one where agents performed model-free-goal learning, and one where they did not[[7]](#footnote-7). We then analyzed the agents’ behavior by the same process as in the behavioral tasks.

Results

In the simulation with model-free-goal learning, on congruent goal trials agents chose the congruent goal 66.3% of the time after a reward and 51.2% of the time after a punishment (t(199) = -.694, p < .0001). The simple mixed-effects model on congruent goal trials estimated a model-free-goal coefficient of .081 (z = 7.35,p < .0001), and was preferred to a null model ((2) = 63.1, p < .0001). The complete mixed-effect model showed similar results.

On incongruent goal trials, agents chose what would have been the congruent goal 48.6% of the time after a reward and 47.6% of the time after a punishment (t(186) = -.292, p = .77). The simple mixed-effect model estimated a model-free-goal coefficient of .0098 (z = 0.649,p = .516), and was not preferred to a null model ((2) = .421, p = .81). The model which combined congruent and incongruent trials showed a significant interaction between model-free-goal value and critical trial type ((3) = 14.4, p < .005).

In the simulation where agents did not perform model-free-goal learning, agents showed no difference in behavior following a reward versus a punishment on congruent goal trials (t(199) = .71, p = .481). The simple mixed-effects model on congruent trials was not preferred to a null model ((2) = .483, p = .786). Neither was the complete model ((4) = 2.33, p = .675). The model combining congruent and incongruent trials did not show a significant interaction effect ((3) = 0, p = 1).

Together, these results show that our experiment detects a model-free-goal effect if and only if agents are actually using model-free-goal learning.

1. If a drift would have caused a reward level to exceed +5 or -4, the reward would ‘rebound’ by however much it would have gone over. For example, if the reward were at +3 and the drift were +5, the next reward would be . [↑](#footnote-ref-1)
2. If the last time a subject got to the color blue was *t1* rounds ago, and she received *r1* points, the model-based value of numbers 1 and 3 would be . If the last time a subject selected number 3 was *t2* rounds ago, and she received *r2* points, the model-based value of number 3 would be . was fixed at .85. [↑](#footnote-ref-2)
3. In both of our mixed-effects models we included all random slopes, and allowed full correlation among all the random slopes and the random intercept. [↑](#footnote-ref-3)
4. Note that congruent goal trials here are equivalent to congruent goal trials in Experiments 1 and 2, because in those experiments all trials were color trials. [↑](#footnote-ref-4)
5. We first estimated a model with the two main effects (type of critical trial and model-free-goal value) and an interaction, but the model was overspecified. Dropping the main effect of critical trial type eliminated that overspecification. Thus, we only present results from the model with the main effect of model-free-goal value and the interaction between model-free-goal value and type of critical trial. [↑](#footnote-ref-5)
6. was sampled from a uniform distribution from 0 to 1, which we denote as U(0,1). was sampled from U(.5,1). was sampled from U(0,1.5). For the weights, three variables – ,,and – were sampled from U(0,1), and then and . [↑](#footnote-ref-6)
7. In the second simulation , forcing agents to use only model-based and model-free mechanisms. [↑](#footnote-ref-7)